

Summarizing Reviews with Variable-length Syntactic Patterns and Topic Models

Trung Nguyen¹ and Alice Oh¹

Computer Science Department, KAIST, Daejeon 305-701, South Korea

Abstract. We present a novel summarization framework for reviews of products and services by selecting informative and concise text segments from the reviews. Our method consists of two major steps. First, we identify five frequently occurring variable-length syntactic patterns and use them to extract candidate segments. Then we use the output of a joint generative sentiment topic model to filter out the non-informative segments. We verify the proposed method with quantitative and qualitative experiments. In a quantitative study, our approach outperforms previous methods in producing informative segments and summaries that capture aspects of products and services as expressed in the user-generated pros and cons lists. Our user study with ninety users resonates with this result: individual segments extracted and filtered by our method are rated as more useful by users compared to previous approaches by users.

1 Introduction

Online reviews of products and services are an important source of knowledge for people to make their purchasing decisions. They contain a wealth of information on various product/service aspects from diverse perspectives of consumers. However, it is a challenge for stakeholders to retrieve useful information from the enormous pool of reviews. Many automatic systems were built to address this challenge including generating aspect-based sentiment summarization of reviews [1,8,19] and comparing and ranking products with regard to their aspects [13]. In this study we focus on the problem of review summarization, which takes as input a set of user reviews for a specific product or service entity and produces a set of representative text excerpts from the reviews.

Most work on summarization so far used sentence as the unit of summary. However, we do not need a complete sentence to understand its main communicative point. Consider the following sentence from review of a coffee maker: ‘My mum bought me this one, and I have to say it makes really awful tasting coffee’. To a buyer looking for an opinion about the coffee maker, only the part ‘makes really awfultasting coffee’ is helpful. Being able to extract such short and meaningful segments from lengthy sentences can bring significant utilities to users. It reduces their reading load as well as presents more readable summaries on devices with limited screen size such as smart phones.

This motivates our main research question of how to extract concise and informative text from reviews of products and services that can be used for

summarization. Previous work has ignored the differences in product and service reviews, which is questionable. To the best of our knowledge, this is the first work that studies and compares summarization for the two domains in details. We propose to extract text segments that match against pre-defined syntactic patterns that occur frequently in reviews of both products and services. However, the extracted segments should be subjected to some selection or filtering procedure as not all matching candidates are likely to contain rich information. Our proposed selection mechanism is based on the observation that segments containing users’ opinions and evaluations about product and service aspects carry valuable information. This motivates the use of output of joint sentiment topic models to discriminate between desirable and non-desirable text segments. Since joint sentiment topic models capture sentiments that are highly associative with aspects, they are well suited for selecting informative segments from the pool of extracted candidates.

The major contributions of our work are as follows.

1. A new joint sentiment-topic model that automatically learns polarities of sentiment lexicons from reviews.
2. Identification of five frequently occurring syntactic patterns for extracting concise segments from reviews of both products and services.
3. Demonstration of the effective application of topic models to select informative variable-length segments for review summarization.
4. Production of summaries that recall important information from review entities’ characteristics.

The rest of the paper is structured as follows. We begin with the related literature in review summarization and joint sentiment topic models in Sect. 2. Next we describe our extension to a topic model and its improvements over previous models in Sect. 3. We then introduce our proposed extraction patterns and procedures for segment selection in Sect. 4. We present our experiments and evaluation in Sect. 5 and 6 and conclude in Sect. 7.

2 Related work

We first look at how text excerpts are extracted from reviews in the existing literature. Previous studies mainly generated aspect-based summary for products and services by aggregating subjective text excerpts related to each aspect. Different forms of the excerpts include sentence [8], concise phrase composing of a modifier and a header term [16], adjective-noun pair extracted based on POS tagging and the term-frequency of the pair [23], and phrase generated by rules [15]. Some limitations of these previous work are i) they only worked with the simplistic adjective-noun pairs or specific form of reviews such as short comments, and ii) experiments were carried out with reviews of services only. Our approach to extract text segments by matching variable-length linguistic patterns overcome these shortcomings and can generalize well for free-text reviews of both products and services.

Various methods for selecting informative text fragments were applied in previous research, such as matching against pre-defined or frequently occurring aspects [1,8], ranking frequency [23], and topic models [17,20,22]. We are interested in the application of joint sentiment topic models as they can infer sentiment words that are closely associative with an aspect. This is an important property of polarity of sentiment words as pointed out in [5,11,13,18], and recently several joint topic models have been proposed to unify the treatment of sentiment and topic (aspect) [9,11,17,21]. Applications of these models have been limited to sentiment classification for reviews, but we hypothesize that they can also be helpful in summarization. We focus our next discussion on previous joint models in comparison to our proposed model.

One of the earliest work is the Topic-Sentiment Model (TSM) [17], which generates a word either from a topic or one of the two additional subtopics – sentiments, but it fails to account for the intimate interplay between a topic/aspect and a sentiment. TSM is based on pLSI whereas more recent work ([9,11,20]) uses or extends Latent Dirichlet Allocation (LDA) [2]. In the Multi-Aspect Sentiment (MAS) model [20], customer ratings are incorporated as signals to guide the formation of pre-defined aspects, which can then be used to extract sentences from reviews that are related to each aspect. In the Joint Sentiment/Topic (JST) model [11], and the Aspect and Sentiment Unification Model (ASUM) [9], each word is assumed to be generated from a distribution jointly defined by a topic and a sentiment (either positive or negative). As a result, JST and ASUM learn words that are commonly associated with an aspect although the models are incapable of distinguishing between sentiment and non-sentiment lexicons. We propose a new model that leverages syntactic information to identify sentiment lexicons and automatically learn their polarities from the co-occurrences of words in a sentence. This allows the model to bootstrap using a minimum set of sentiment seed words, thereby alleviating the need for information that is expensive to obtain such as ratings of users for reviews [20] or large lists of sentiment lexicons [11].

3 A Topic Model for Learning Polarity of Sentiment Lexicons

Our key modelling assumption for reviews is that a sentence expresses an opinion toward an aspect via its sentiment component. For example, in the sentence ‘The service was excellent’, only the word ‘excellent’ carries the positive sentiment. This is not a new assumption as adjectives and adverbs are commonly considered the main source of sentiment in a sentence in existing literature. Our model leverages on this type of knowledge to locate sentiment words in a sentence with relatively high confidence.

3.1 Generative Process

The formal generative process of our model for the graphical representation in Fig. 1 is as follows (see Table 1 for the list of notations).

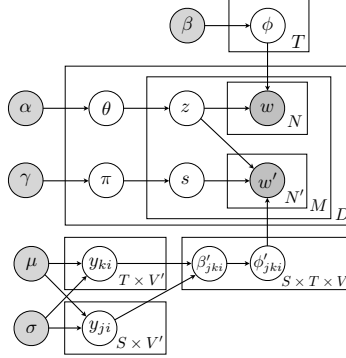


Fig. 1. Graphical representation of the model.

- For every aspect k , draw a distribution of non-sentiment words, $\phi_k \sim \text{Dir}(\beta)$, and two distributions of sentiment words, $\phi'_{jk} \sim \text{Dir}(\beta'_{jk})$, where $j = 0$ denotes positive polarity and $j = 1$ denotes negative polarity.
- For each review d ,
 - ◊ Draw a sentiment distribution $\pi_d \sim \text{Dir}(\gamma)$
 - ◊ Draw a topic distribution $\theta_d \sim \text{Dir}(\alpha)$
 - ◊ For each sentence c in document d ,
 - Choose a topic $z = k \sim \text{Mult}(\theta_d)$ and a sentiment $s = j \sim \text{Mult}(\pi_d)$
 - Choose words $w \sim \text{Mult}(\phi_k)$ to discuss aspect k and sentiment words $w' \sim \text{Mult}(\phi'_{jk})$ to convey the sentiment j toward k .

Notice in the graphical model that the part of a sentence which emanates the sentiment is observed. In our implementation, we treat all adjectives and adverbs as w' and remaining words as w in the generative procedure, but this is not a restriction imposed on the model. It is easy to incorporate prior knowledge about words that convey sentiment into the model. For example, we can instruct the model that words such as *love*, *hate*, *enjoy*, *worth*, *disappoint* are sentiment words, even though they are not adjective nor adverb.

Our main extension deals with the word smoother β' for sentiment words. Each sentiment word i is associated with a topic dependent smoothing coefficient y_{ki} for topic k and a sentiment dependent smoothing coefficient y_{ji} for sentiment j . We then impose that

$$\beta'_{jki} = \exp(y_{ki} + y_{ji}), \quad y_{ki} \sim N(0, \sigma_1^2), \quad y_{ji} \sim N(0, \sigma_2^2). \quad (1)$$

This modeling allows us to incorporate polarity of sentiment words as side information. The polarity of sentiment lexicon i in a corpus is represented by the values of y_{ji} ; this is to assume that the polarity of i is its intrinsic property as the corpus is about a specific domain [3]. The topic dependent smoother y_{ki} is introduced to accommodate the different frequency of association between the sentiment word i and different aspects.

Table 1. List of notations used in the paper (senti = sentiment, dist. = distribution)

d, c, w, w', k, j	: review, sentence, non-senti word, senti word, topic/aspect, sentiment
T, S, V, V'	: number of topics, sentiments, non-senti words, senti words
π_d, θ_d	: sentiment distribution, topic distribution of the review d
ϕ_k, ϕ'_{jk}	: word dist. of topic k , senti word dist. of topic k and senti j
y_{ji}	: polarity of senti word i with sentiment j
y_{ki}	: smoother for dependency between topic k and senti word i
β'_{jki}	: word smoother for senti word i with topic k and senti j
$\alpha, \beta, \gamma, \mu, \sigma$: hyperparameters
$n_{ki}^{TW}, n_{jki}^{STW}$: counts of word i being assigned topic k , senti word i being assigned topic k and senti j
n_{dk}^{DT}, n_{dj}^{DS}	: counts of sentences in d being assigned topic k , sentences in d being assigned senti j

3.2 Inference

In order to perform inference we alternate between two procedures: sampling and maximum a posteriori. The sampler assigns values for the latent variables: the topics and sentiments of sentences. Using a collapsed Gibbs sampler [7], new values for the topic and sentiment of a sentence c in document d are drawn from the conditional probability

$$p(z_{dc} = k, s_{dc} = j | \text{rest}) \propto \frac{\prod_{i \in A(dc)} (n_{\setminus ki}^{TW} + \beta)}{\prod_{x=0}^{|A(dc)|-1} \sum_{i=1}^V n_{\setminus ki}^{TW} + V\beta + x} \frac{\prod_{i \in S(dc)} (n_{\setminus jki}^{STW} + \beta'_{jki})}{\prod_{x=0}^{|S(dc)|-1} \sum_{i=1}^{V'} (n_{\setminus jki}^{STW} + \beta'_{jki}) + x} (n_{\setminus dk}^{DT} + \alpha)(n_{\setminus dj}^{DS} + \gamma) \quad (2)$$

where $S(dc)$ is the set of sentiment words in c and $A(dc)$ is the set of remaining words. The ' \setminus ' notation means not counting the sentence being sampled.

We estimate the value for β' and y from a maximum a posteriori procedure, optimizing β' over y and the assigned values of the latent variables. The negative log prior is

$$-\log p(\beta') = S \sum_{k,i} y_{ki} + T \sum_{j,i} y_{ji} + \sum_{j,k,i} \frac{(y_{ki} + y_{ji})^2}{2\sigma^2} \quad (3)$$

where $\sigma^2 = \sigma_1^2 + \sigma_2^2$. The collapsed negative log likelihood (dependent on sentiment words only) is

$$L_{\beta'} = \sum_{j,k} [\log \Gamma(\bar{n}_{jk} + \bar{\beta}'_{jk}) - \log \Gamma(\bar{\beta}'_{jki})] + \sum_{j,k,i} [\log \Gamma(\beta'_{jki}) - \log \Gamma(n_{jki}^{STW} + \beta'_{jki})] \quad (4)$$

where $\bar{n}_{jk} = \sum_{i=1}^{V'} n_{jki}^{STW}$, $\bar{\beta}'_{jk} = \sum_{i=1}^{V'} \beta'_{jki}$, and Γ is the Gamma function. We use the L-BFGS optimizer [14] to minimize the objective function $L_{\beta'} - \log p(\beta')$ by taking its partial derivatives with respect to y_{ki} and y_{ji} .

A sample from the Markov chain in the sampler can be used to estimate the distributions of interest. The approximate probabilities of sentiment j in document d ($\hat{\pi}_{dj}$), topic k in document d ($\hat{\theta}_{dk}$), non-sentiment word i in topic k ($\hat{\phi}_{ki}$), and sentiment word i in topic k and sentiment j ($\hat{\phi}'_{jki}$) are

$$\begin{aligned}\hat{\pi}_{dj} &= \frac{n_{dj}^{DS} + \gamma}{\sum_{j'=1}^S n_{dj'}^{DS} + S\gamma} , & \hat{\theta}_{dk} &= \frac{n_{dk}^{DT} + \alpha}{\sum_{k'=1}^T n_{dk'}^{DT} + T\alpha} , \\ \hat{\phi}_{ki} &= \frac{n_{ki}^{TW} + \beta}{\sum_{i'=1}^V n_{ki'}^{TW} + V\beta} , & \hat{\phi}'_{jki} &= \frac{n_{jki}^{STW} + \beta'_{jki}}{\sum_{i'=1}^{V'} (n_{jki'}^{STW} + \beta'_{jki'})} .\end{aligned}\quad (5)$$

3.3 Aspect and Sentiment Classification Using Output of the Model

As stated in the introduction, we attempt to use the outputs of this model to improve the selection of informative segments for summarization. We define the topic classifier of an arbitrary segment of n words $G = (w_1, w_2, \dots, w_n)$ as

$$\arg \max_k p(k|G) = \arg \max_k \sum_{i=w_1}^{w_n} (\log \hat{\phi}_{ki} + \sum_j \log \hat{\phi}'_{jki}) . \quad (6)$$

To classify the sentiment of a segment G , we use the sentiment value y_{ji} learned from the model. We define the polarity of G as

$$polarity(G) := \sum_{\text{sentiment word } i \in G} polarity(i) = \sum_{\text{sentiment word } i \in G} y_{0i} - y_{1i} . \quad (7)$$

G is classified as positive if $polarity(G) \geq 0$ and as negative if $polarity(G) < 0$.

4 Summarization Using Syntactic Patterns and Topic Models

In this section we present our framework for variable-length segment-based summarization of reviews. We first describe the five frequently occurring syntactic patterns in reviews that are used to extract candidate text segments. We then discuss the use of topic models in selecting meaningful segments from the set of extracted candidates. We also present an independent framework for evaluation of the summaries comprising segments regardless of the approaches.

4.1 Extraction Patterns

Central to our summarization system is how to extract meaningful, informative text segments out of a sentence. We use sentence syntax to guide the extraction process by defining patterns of lexical classes for matching against text segments. The purpose is to extract semantically meaningful unit of text in a sentence that can be understood without extra context. In the particular task of summarizing

reviews for products and services, we want to capture units that contain sentiments toward aspects. This type of segments is important because it expresses and formulates opinions about the entity being reviewed.

Based on the above observation, we identify five most common extraction patterns to capture a variety of text segments in both product and service reviews as follows. First we use POS tagger to tag all pros and cons items available in our data sets of restaurant and coffee maker reviews (see Sect. 5.1). The pros and cons are relatively short and meaningful, and can therefore be suitable representatives of the text segments that we want to generate. The resulting sequences of tags are then ranked based on their frequency. After carefully studying the top ranked patterns we select the five most productive ones listed in Table 2.

Table 2. Extraction patterns and their occurrences in data sets

no.	the pattern	example	restaur	coff makers
1	nn? vb dt? rb* jj nn	instruction booklet includes clear instruction	56468	23210
2	nn? vb rb* jj to vb	filter basket is simple to remove	4226	3770
3	nn? vb rb* jj	design is striking, tasted fresh	130853	30449
4	rb* jj to vb nn?	easy to clean, wide enough to insert a K-Cup	5937	5288
5	rb* jj nn	very good food, most expensive pod brewer	197123	69273

We use the same notations as in regular expression, where the constituent parts correspond to lexical categories as specified by the PennTree bank. For simplicity, a single tag is used to represent different forms of a category; i.e., **jj** represents adjective and matches all of JJ, JJR and JJS. Also, **nn** matches a noun phrase rather than just a single word. We further restrict that each segment must match the longest pattern. This means, for example, a segment matching pattern 1 in a sentence is consumed and no longer available for matching pattern 5. Each pattern also has its negation form easily constructed from its positive form, hence we do not show in the table.

4.2 Selecting Informative Segments using Topic Models

Candidate segments can be meaningless even if they match the defined extraction patterns. For example, ‘final thought’ and ‘several hour’ are instances of pattern 5, but they reveal no interesting information. Furthermore, the sheer number of text segments matching the patterns (Table 2) requires us to be selective in finding segments to include in summaries.

We observe that informative segments often contain words that convey opinions about aspects of entities. Since the aspect-sentiment intimate interplay is modeled and learned by our joint sentiment-topic model, we propose the following filters to prune less informative segments using the output the model.

Baseline No filtering, i.e., keep all matching segments.

- AW** Eliminate a segment if it does not contain one of the top X most probable words of the segment’s inferred aspect.
- SW** Eliminate a segment if it does not contain one of the top Y most probable sentiment words of the segment’s inferred sentiment and aspect.
- RANK** Rank all segments having the same inferred sentiment and aspect in order of their probabilities and eliminate the bottom half segments.

It is possible to use previous joint sentiment topic models, such as ASUM [9] and JST [11], for the filtering purpose. Note that ASUM and JST output word distributions for each pair of sentiment and aspect; hence, **ASUM** and **JST** are in effect both sentiment classifier and filter:

- ASUM** Eliminate a segment if it does not contain one of the top Z most probable words of the segment’s inferred senti-aspect using the ASUM model.
- JST** Same as **ASUM** except that JST is used in placed of ASUM.

A complete procedure for summarization would need a sentiment classifier component for segments as sentiment-based summaries are preferred by users [10]. In addition to our model-based sentiment classifier, we introduce another sentiment classifier based on SentiWordNet (SWN) [4], a popular lexical resource for opinion mining, using the same approach as in [5]. For convenience, we call our model-based classifier **SEN** and the SWN-based classifier **SWN**.

Various procedures for retaining quality segments can then be constructed by combining different sentiment classifiers and filters. For example, we may first use **SEN** to classify sentiment of a segment and then use both **AW** and **SW** to discard non-qualified segments. We name such procedure **SEN+AW+SW**, with the convention that the output of a preceding classifier/filter is the input to the next classifier/filter whenever applicable.

4.3 A Framework for Segment-based Summary Evaluation

We now introduce a framework for automatically evaluating the extraction patterns at the levels of segment and entity (a specific product or service). This framework is independent of the way segments are generated and therefore can be applied to any method that uses segment as the unit of summary.

Each entity E has a candidate summary $E^C = \{Y | Y \text{ matches one of the patterns}\}$ and a reference summary $E^R = \{X | X \text{ is in the gold standard summary of } E\}$. For $Y \in E^C$ and $X \in E^R$, we measure the similarity of their content using precision and recall

$$P(X, Y) := \frac{\text{skip2}(X, Y)}{\binom{|Y|}{2}}, \quad R(X, Y) := \frac{\text{skip2}(X, Y)}{\binom{|X|}{2}}$$

where $\text{skip2}(X, Y)$ is the number of skip-bigram matches between X and Y (termed ROUGE-SU in [12]). For a candidate segment $Y \in E_C$, define $R(Y) = R(X_{max}, Y)$ and $P(Y) = P(X_{max}, Y)$ where $X_{max} = \arg \max_{X \in E^R} R(X, Y)$.

For an entity E , the average precision $P_{skip}(E) = \sum_{Y \in E^C} P(Y) / |E^C|$ and recall $R_{skip}(E) = \sum_{Y \in E^C} R(Y) / |E^C|$ tells us how similar the content of extracted

segments is to a reference set of segments on average. We also want to assess how many portion of the reference summary is recovered and what percentage of the candidate summary is useful. For this reason, we introduce $P(E)$ and $R(E)$ to measure the precision and recall for the candidate summary set E^C of E :

$$P(E) := \frac{\sum_{Y \in E^C} \mathbf{1}_A\{\mathbf{Y}\}}{|E^C|}, \quad R(E) := \frac{\sum_{X \in E^R} \mathbf{1}_B\{\mathbf{X}\}}{|E^R|} \quad (8)$$

where $\mathbf{1}_A\{\mathbf{Y}\}$ and $\mathbf{1}_B\{\mathbf{X}\}$ are indicator functions; $\mathbf{A} = \{Y \mid R(Y) \geq \alpha\}$ and $\mathbf{B} = \{X \mid \exists Y \in \mathbf{A} \text{ s.t. } R(X, Y) = R(Y)\}$ where α is a recall threshold for a candidate summary to be considered useful.

A good measure for a reference summary of an entity E must be a combination of the segment-level recall (precision), $R_{skip}(E)$ and the entity-level recall (precision), $R(E)$. A simple combination is the average of the two, i.e., $R_{cb}(E) = (R_{skip}(E) + R(E))/2$ and $P_{cb}(E) = (P_{skip}(E) + P(E))/2$.

Since we typically work with data that contains a large set of review entities, it is convenient to report the results using the following summarization statistics:

$$P_s = \frac{\sum_{i=1}^N \sum_{Y \in E_i^C} P(Y)}{\sum_{i=1}^N |E_i^C|}, \quad P_e = \sum_{i=1}^N P(E_i)/N, \quad R_e = \sum_{i=1}^N R(E_i)/N,$$

$$R_s = \frac{\sum_{i=1}^N \sum_{Y \in E_i^C} R(Y)}{\sum_{i=1}^N |E_i^C|}, \quad P = \sum_{i=1}^N P_{cb}(E_i)/N, \quad R = \sum_{i=1}^N R_{cb}(E_i)/N.$$

5 Experiments

We experimented using reviews of coffee makers as representative for the product domain and reviews of restaurants as representative for the service domain. We describe our data sets and experimental set-ups in 5.1. In 5.2 we give example of the topics and sentiment words learned by the model. We analyze the effectiveness of extraction patterns in 5.3 and compare the performance of different sentiment classifiers and segments filters in 5.4.

5.1 Data Sets and Experimental Set-ups

Our data sets consist of restaurant reviews and coffee maker reviews. For each review, we collected its free-format text content and its pros and cons lists if available.

- **RESTAURANTS** 50,000 reviews of 5,532 restaurants collected from Citysearch New York. This data is provided by Ganu, et al. [6].
- **COFFEEMAKERS** 23,411 reviews of 534 coffee makers collected from epinions.com.

Our first step is to fit the joint sentiment topic model to each data set. Data is pre-processed as in other standard topic models, in which sentences are tokenized

by the punctuations: ‘.’, ‘!’, and ‘?’ . The hyperparameters are set as $\alpha = 0.1$, $\beta = 0.01$, $\gamma = 0.1$ for both positive and negative sentiment; the number of aspects is 7 for both corpora.

We incorporated prior sentiment information into the model using sentiment seed words in analogy to [9]. After running the sampler for a burnin period of 500 iterations, we interleaved it with the optimizer, optimizing over y_{ki} and y_{ji} every 100th step of sampling. We trained the model in 2000 iterations for both data sets and used the last sample in the chain in all of our experiments.

In the segments selection step, the maximum number of words in a sequence is set to 7 and the number of top words for **AW**, **SW**, **ASUM**, and **JST** is set to 200, 100, 300, and 300, respectively. We used a value of 0.25 for the recall threshold α in Eq. 8. All parameters were set empirically after many experiments.

In order to evaluate the quality of segments and summaries using the framework in 4.3, a reference summary must be obtained for each review entity. We aggregate the pros written by all reviewers for an entity as its pros gold standard and similarly for its cons standard (duplicated entries are removed). To construct an entity’s candidate summaries, the procedures in 4.2 are applied to the segments extracted from all of its reviews. The sentiment classifier in a procedure partitions the entity’s segments into a positive candidate summary and a negative candidate summary. The candidates are evaluated against their counterpart reference summaries independently.

Table 3. Example inferred topics (restaurants: row 1-3, coffee makers: row 4-6)

top aspect words	top positive words	top negative words
sauc, chicken, chees, salad, shrimp, soup, fri, potato, rice	good, delici, best, great, fresh, love, perfect, excel, amaz, tasti	dri, disappoint, tasteless, cold, soggi, bad, fri, rare, medium
music, place, bar, decor, room, table, wall, seat, atmosphere	great, nice, good, love, beauti, enjoy, romant, perfect, friend,	loud, noisi, bad, littl, small, crowd, dark, expans, back
wait, table, waiter, seat, minut, reserv, order, ask, told, manag	friend, nice, worth, great, at-tent, prompt, long, enjoy, quick	rude, bad, wrong, final, empti, horribl, terribl, poor, worst
coffe, bean, cup, grind, ground, brew, grinder, espresso, tast	good, like, fresh, great, best, hot, strong, fine, french, perfect	weak, bad, disappoint, wast, grind, wrong, unfortun, bitter
filter, clean, basket, water, pa-per, rins, dishwash, gold, use	easi, clean, perman, like, remov, easili, good, recommend, safe	difficult, clean, wet, bad, im-poss, perman, wast, not easi
servic, game, custom, warranti, repair, ship, product, send, call	good, new, back, great, free, thank, well, happi, local, origin,	back, disappoint, poor, bad, wrong, negative, defect, sorri

5.2 Topics and Polarities of Sentiment Words Learned by the Model

Example of topics inferred by the model is given in Table 3. Each topic has three distributions where one distribution (first column) consists descriptive words about the aspect and two distributions (remaining columns) consist evaluative

words directing the aspect. Except the common sentiment words such as *good*, *great*, *bad*, *wrong* that are associated with most aspects due to their frequent usage, positive and negative sentiment lexicons look highly related to their corresponding aspects. For example, the model discovers that people are more likely to praise the food with *delicious*, *best*, *fresh*, and *tasty* and disapprove food that is *dry*, *tasteless*, *cold* or *soggy*. Such results can be very helpful for the exploratory purpose of understanding what aspects reviewers care and comment about.

Table 4 demonstrates the effectiveness of our model in learning the polarities of domain-specific sentiment lexicons (the seed words used for bootstrapping are excluded). To verify this claim we compare with the **SWN** classifier described in 4.2 in a classification task for noun phrases. SWN leverages synsets in WordNet and so, in some sense, it captures the context-dependent sentiment of a word. We used a set of 929 positive and 236 negative noun phrases obtained from an external set of restaurant reviews in [5]. All phrases are unique and manually annotated with their true sentiments. Our classifier outperforms **SWN** in classification accuracy for both the positive (90.1% vs. 83.4%) and negative (74.6% vs. 66.1%) categories. This shows that our model is quite accurate in assigning sentiment score to domain-specific lexicons compared to the more general propagation approach in SWN.

Table 4. Selected lexicons with their sentiment polarities

positive lexicons in restaurant reviews
knowledgeable, helpful, unique, courteous, prompt, cozy, terrific, wonderful, affordable, superb, warm, impeccable, outstanding, elegant, consistent, fabulous, charming
negative lexicons in restaurant reviews
tasteless, mediocre, bland, inedible, dry, ridiculous, lousy, overpriced, flavorless, average, unacceptable, obnoxious, soggy, bare, bore, tough, unfriendly, horrendous, stale
positive lexicons in coffee maker reviews
simple, ready, fresh, correct, removable, automatic, impressive, stainless, large, free, light, strong, rich, reasonable, amazing, fast, clear, wonderful, delicious, quick, sturdy
negative lexicons in coffee maker reviews
difficult, impossible, inferior, loud, lousy, dull, defective, stupid, sticky, dirty, faulty, uneven, weak, noisy, stiff, frustrating, dissatisfied, smelly, unclear, erratic, leak, slow

5.3 Evaluation of Extraction Patterns

We now analyze how different extraction patterns behave when applied to the service domain and the product domain (Table 5 and 6). We use **AW+SEN+SW** procedure because it produced the best result among all methods. For reviews of restaurants, pattern 3 and 5 are the most productive with superior average precision and recall at both segment-level and entity-level compared to the rest. They account for more than half of an entity’s pros and cons reference. This is probably due to the prevalence of sentences such as ‘the service was good’

in restaurant reviews. The result is consistent with the current literature where adjectives and nouns are commonly used to detect sentiments and aspects in reviews for services. It is worth noticing that extracting any thing other than adjective-noun pairs may degrade the quality of summarization as the scores for pattern 2 and 4 are overwhelmingly low.

Table 5. Comparison of extraction patterns for services.

patt	pros				cons			
	P _s	R _s	P _e	R _e	P _s	R _s	P _e	R _e
1	20.2	51.1	74.4	30.4	14.6	39.6	55.7	14.8
2	23.7	26.4	54.6	1.9	7.9	22.4	63.6	0.5
3	31.8	65.7	83.3	40.6	26.4	57.6	70.1	21.5
4	21.3	28.8	56.1	2.6	6.3	15.5	35.7	0.37
5	25.9	53.5	72.7	49.7	18.1	37.8	52.7	31.2

The behaviors of extraction patterns are trickier for the product domain as can be seen in Table 6. There is no dominating pattern in terms of high precision and recall at both segment and entity level. In particular, pattern 3 and 5 still recover a large portion of an entity’s reference summary; however, the average quality of their matching segments (R_s) is the lowest among all patterns. Pattern 2 and 4 perform badly when used with the service domain but are more useful in the product domain, producing the highest quality segments ($R_s = 66.3$ and 69.4 for positive; $R_s = 48.6$ and 43.8 for negative). Although they do not appear as frequently in reviews as other patterns, they tend to carry more meaning in their words that it is hard to ignore them. Hence, all five patterns can contribute to the extraction of informative segments for summarization. This shows that doing summarization for products is harder than for services; and, care should be exercised when generalizing results from one domain to the other.

Table 6. Comparison of extraction patterns for products.

patt	pros				cons			
	P _s	R _s	P _e	R _e	P _s	R _s	P _e	R _e
1	22.1	59.6	70.8	30.1	19.5	44.2	59.7	17.5
2	45.5	66.3	78.7	6.6	30.2	48.6	64.3	2.9
3	33.2	54.4	65.9	26.7	27.0	37.7	51.1	15.0
4	52.7	69.4	78.5	8.6	30.9	43.8	61.3	3.5
5	26.0	50.8	59.6	38.9	26.2	40.8	56.3	25.0

5.4 Evaluation of Sentiment Classifiers and Segment Filters

Results in the previous section suggest to use different syntactic patterns for summarization of the service and product domains. We used patterns 1, 3, and

5 for services and all patterns for products in all of our experiments in this section.

We applied seven different procedures for selecting candidate segments to compare the effects of 2 sentiment classifiers (**SEN** and **SWN**) and 5 filters (**AW**, **SW**, **RANK**, **ASUM**, and **JST**). The results are depicted in Table 7 and 8. The good overall performance of the **Baseline+SWN** procedure in both domains indicates that the proposed patterns extract good segments for summarization.

Table 7. Comparison of classifiers and filters for services.

procedure	pros						cons					
	P _s	R _s	P _e	R _e	P	R	P _s	R _s	P _e	R _e	P	R
Baseline+SWN	24.4	48.6	64.8	65.4	44.5	56.7	17.6	29.0	42.1	45.0	29.7	36.8
AW+SWN	24.8	55.5	71.4	60.4	47.9	57.6	19.0	33.8	47.6	37.8	33.1	35.5
AW+SEN	23.8	52.3	67.5	62.8	45.5	57.2	21.3	47.6	61.5	36.9	41.3	42.1
AW+SEN+RANK	29.3	52.6	66.7	44.5	47.8	48.2	26.0	46.8	60.4	23.1	43.0	34.8
AW+SEN+SW	25.8	56.4	73.3	59.9	49.3	57.8	25.4	58.2	73.5	30.2	49.4	44.0
ASUM	27.4	49.7	66.7	65.8	46.8	57.4	22.0	33.5	47.0	47.4	34.3	40.2
JST	24.6	44.7	59.8	60.4	42.0	51.9	21.3	37.3	49.4	52.3	35.2	43.9

Table 8. Comparison of classifiers and filters for products.

procedure	pros						cons					
	P _s	R _s	P _e	R _e	P	R	P _s	R _s	P _e	R _e	P	R
Baseline+SWN	23.1	49.2	60.9	44.8	41.1	43.9	23.9	34.5	47.3	32.6	34.6	30.7
AW+SWN	21.8	52.6	65.4	42.7	42.8	44.7	23.4	39.0	52.7	30.8	37.6	31.6
AW+SEN	23.9	56.3	68.5	47.0	45.3	48.2	26.1	47.5	62.9	24.7	43.5	32.1
AW+SEN+RANK	29.7	51.2	61.0	32.4	44.0	38.4	32.6	37.3	48.9	21.2	39.4	25.8
AW+SEN+SW	25.5	59.8	71.4	43.8	47.4	48.2	24.3	51.4	65.4	16.7	44.2	29.4
ASUM	25.3	54.8	68.2	49.3	45.7	48.8	26.9	37.8	49.0	28.7	36.7	29.8
JST	27.1	52.0	64.3	44.9	44.7	45.2	25.9	35.7	46.4	28.8	34.5	28.5

Comparing **AW+SWN** and **AW+SEN**, we see that **SEN** is better than **SWN** at sentiment classification. This result agrees with previous section, again confirming the effectiveness of our model in learning sentiments of domain-specific lexicons. **AW+SWN** performs better than **Baseline+SWN**, suggesting that top aspect words can be used to identify more informative segments. The best procedure is **AW+SEN+SW**, which tops all other procedures especially in the cons case. It favors segments that contain common aspect-related words and its associated sentiment lexicons, which are likely to be predominant in the pros and cons lists. ASUM has similar modelling assumption as ours and so it also produces relatively good results. However, the ability of our model to optimize sentiment polarities creates the improvement in performance. **JST** is even inferior to **Baseline+SWN** for half of the cases. This is not surprising given that the JST model is not intended for sentiment lexicons discovery; in contrast, it requires a large list of sentiment seed words to function well. Fi-

nally, **AW+SEN+RANK** always has highest precision for segments but many segments are eliminated and that hurts its performance.

6 Qualitative Evaluation

In this section we complement our results in previous section by qualitatively evaluating the quality of extracted segments with a user study and present example of summaries generated by our approach.

6.1 Quality of Extracted Segments

We carried out an user study with 130 workers from the Amazon Mechanical Turk service. We randomly selected 123 short passages each has 4 to 6 sentences from reviews of coffee makers. The user’s task is to read a passage and rate each text item as ‘very useful’, ‘useful’, ‘somewhat useful’, or ‘useless’ with reference to the passage. We included two types of items for each passage: segments extracted by our approach using the **AW** filter and adjective-noun phrases extracted using tagging and term-frequency as in [23]. Each user performs 6 tasks in which half of them are repetitions of others, thereby allowing us to detect users that give inconsistent ratings. We discarded users who completed their tasks in less than 90 seconds or rated half of the items inconsistently. Of the remaining 90 qualified users, 13 have not used coffee makers before whereas 60 have used for more than two months.

In total there were 358 unique segments, each rated 5.3 times and 470 unique word pairs, each rated 5.7 times. We converted the ratings into a numeric scale from 4 to 1 with 4 being ‘very useful’ and 1 being ‘useless’. On average, users rated the segments extracted by our method as 3.01 compared to 2.53 for the adjective-noun phrases. The higher rating is not merely due to segments having more words, as we observed that users typically give an adjective-noun word pair a same or higher rating than a segment if the two carry the same message. For example, ‘carafe stays hot’ and ‘hot carafe’ are same but the former has a rating of 2.7 whereas the latter has a rating of 3.1. Therefore the higher average rating for segments is a strong evidence that they convey more valuable information than adjective-noun word pairs.

Table 9 elaborates further on this evidence by showing example of the segments and phrases rated as very useful by users. As can be seen, the segments are quite complete semantically whereas the phrases can be rather short in their meaning, which may require interpretation from users.

6.2 Example Summaries

Below we show examples of a restaurant review and a coffee maker review together with the segments extracted as their summaries.

Review of restaurant: *The space is small but cozy, and the staff is friendly and knowledgeable. There was some great music playing, which kind of made me*

Table 9. Example of highly rated segments and phrases

segment
very easy instruction, almost completely unscrewed to pour, buttons are easy to press, cup is always fresh, coffee pot is very hard to take, closing is easy, makes really awful tasting coffee, feature works fine, machine brews a great cup every time, machine is very simple to use, machine is programmable, carafe is dripless
adjective-noun phrase
affordable maker, better tasting, correct time, filtered water, finished quality, difficult place, fresh tasting, good customer, good tasting, new recipe, removable basket, optimal temperature, cheap use, easy closing, darker flavor, hot cup, great pot

feel like I was on vacation some place far away from Astoria. There are a lot of really great vegetarian options, as well as several authentic Turkish dishes. If you're still wasting time reading this review, stop now and head straight for Mundo. Your stomach could already be filled with tons of deliciousness.

Summary: staff is friendly, space is small, some great music playing , several authentic Turkish dishes, really great vegetarian options.

Review of coffee maker: *I bought this machine about a week ago. I did not know which machine in the store to get, but the sales clerk helped me make the decision to buy this one. It is incredibly simple to use and the espresso is great. The crema is perfect too. My latte's rival those in coffee houses and I am saving a ton of money. The "capsules" must be ordered from the Nespresso website, but they are usually at your door in 48 hours via UPS...*

Summary: incredibly simple to use, espresso is great, crema is perfect.

In both cases the summaries express the gist of each review relatively well. Looking at the sentence where a segment is extracted from, it can be seen that the segment conveys the main talking point of the sentence. Additionally, each segment does express an opinion about some aspect of the coffee maker or the restaurant. Recall that our key assumption in modeling reviews is that each sentence has a sentiment and an aspect. Therefore extracting segments the way we propose is likely to capture the main content of a sentence.

7 Conclusions

In this paper we have describe a framework for extracting and selecting informative segments for review summarization of products and services. We extract candidate segments by matching against variable-length syntactic patterns and select the segments that contain top sentiment and aspect words learned by topic models. We proposed a new joint sentiment topic model that learns the polarity of aspect dependent sentiment lexicons. Qualitative and quantitative experiments verify that our model outperforms previous approaches in improving the quality of the extracted segments as well as the generated summaries.

References

1. S. Blair-Goldensohn, K. Hannan, R. McDonald, T. Neylon, G. A. Reis, and J. Reynar. Building a sentiment summarizer for local service reviews. In *WWW '08*.
2. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.
3. Y. Choi, Y. Kim, and S.-H. Myaeng. Domain-specific sentiment analysis using contextual feature generation. In *CIKM'09*, pages 37–44.
4. A. Esuli and F. Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *LREC06*, pages 417–422.
5. A. Fahrni and M. Klenner. Old wine or warm beer: Target-specific sentiment analysis of adjectives. *Computational Linguistics*, 2(3):60–63, 2008.
6. G. Ganu, N. Elhadad, and A. Marian. Beyond the stars: Improving rating predictions using review text content. In *WebDB*, 2009.
7. T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.
8. M. Hu and B. Liu. Mining and summarizing customer reviews. In *KDD'04*, pages 168–177.
9. Y. Jo and A. H. Oh. Aspect and sentiment unification model for online review analysis. In *WSDM'11*, pages 815–824.
10. K. Lerman, S. Blair-Goldensohn, and R. McDonald. Sentiment summarization: evaluating and learning user preferences. In *EACL'09*, pages 514–522.
11. C. Lin and Y. He. Joint sentiment/topic model for sentiment analysis. In *CIKM'09*, pages 375–384.
12. C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In S. S. Marie-Francine Moens, editor, *ACL'04 Workshop*, pages 74–81.
13. B. Liu. Sentiment analysis and subjectivity. In *Handbook of Natural Language Processing, Second Edition*. CRC Press, 2010.
14. D. C. Liu, J. Nocedal, D. C. Liu, and J. Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical Programming*, 45:503–528, 1989.
15. J. Liu and S. Seneff. Review sentiment scoring via a parse-and-paraphrase paradigm. In *EMNLP'09*, pages 161–169.
16. Y. Lu, C. Zhai, and N. Sundaresan. Rated aspect summarization of short comments. In *WWW'09*, pages 131–140, 2009.
17. Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai. Topic sentiment mixture: modeling facets and opinions in weblogs. In *WWW '07*, pages 171–180.
18. B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *EMNLP '02*, pages 79–86.
19. A.-M. Popescu and O. Etzioni. Extracting product features and opinions from reviews. In *HLT '05*, pages 339–346.
20. I. Titov and R. McDonald. A joint model of text and aspect ratings for sentiment summarization. In *ACL-08:HLT*, pages 308–316.
21. I. Titov and R. McDonald. Modeling online reviews with multi-grain topic models. In *WWW '08*, pages 111–120.
22. X. Xu, T. Meng, and X. Cheng. Aspect-based extractive summarization of online reviews. In *SAC '11*, pages 968–975.
23. K. Yatani, M. Novati, A. Trusty, and K. N. Truong. Review spotlight: a user interface for summarizing user-generated reviews using adjective-noun word pairs. In *CHI '11*, pages 1541–1550.